# An overview of view-based 2D/3D indexing methods

Raluca-Diana Petre[(1)], Titus Zaharia[(1)], Françoise Prêteux[(2)]
[(1)]Institut TELECOM; TELECOM SudParis, ARTEMIS Department; UMR CNRS 8145 MAP5
9 rue Charles Fourier, 91011 Evry Cedex, France
[(2)] Mines ParisTech; 60 Boulevard Saint-Michel, 75272 Paris, France
{Raluca-Diana.Petre, Titus.Zaharia}@it-sudparis.eu, Francoise.Preteux@mines-paristech.fr

## ABSTRACT

This paper proposes a comprehensive overview of state of the art 2D/3D, view-based indexing methods. The principle of 2D/3D indexing methods consists of describing 3D models by means of a set of 2D shape descriptors, associated with a set of corresponding 2D views (under the assumption of a given projection model). Notably, such an approach makes it possible to identify 3D objects of interest from 2D images/videos. An experimental evaluation is also proposed, in order to examine the influence of the number of views and of the associated viewing angle selection strategies on the retrieval results. Experiments concern both 3D model retrieval and image recognition from a single view. Results obtained show promising performances, with recognition rates from a single view higher then 66%, which opens interesting perspectives in terms of semantic metadata extraction from still images/videos.

**Keywords:** indexing and retrieval, 2D and 3D shape descriptors, multiview matching, similarity measures, MPEG-7 standard, 3D meshes.

## 1. INTRODUCTION

The domain of 3D graphics has known a spectacular expansion during the last decade, due to the development of both hardware and software resources. Within this context, more and more industrial applications involving 3D objects have been entered our daily life. Computer aided design (CAD), gaming, special effects and film production, biology, medicine, chemistry, archaeology or geography are some examples of application domains that use intensively 3D object representations. In addition, the amount of available 3D models is continuously increasing, due to the availability of advanced 3D content creation platforms (*e.g*. Maya, 3DSMax, Blender) as well as of low-cost 3D scanners.

The availability of large 3D model repositories throws to the scientific community new challenges in terms of 3D content processing, analysis and representation. One fundamental issue concerns the content re-use within audio-visual production chains. The goal in this case is to retrieve from large 3D databases similar 3D models that can be exploited within the framework of new production projects. Content-based similarity retrieval of 3D mesh models offers a promising solution to this issue. The principle consists of describing in a salient manner the object's visual features, such as shape, colour, texture, motion and then to use them for comparison purposes and automatic 3D object retrieval. Let us note that in most of the cases the key feature exploited for retrieval purposes is the 3D shape, which is the sole available in all cases: 3D models are not necessarily textured, coloured or animated, but they always define a 3D shape.

As a consequence, numerous 3D shape retrieval approaches have been proposed during the last decade. For some comprehensive overviews, the reader is invited to refer to [11, 18, 14, 21].

Among the various families of 3D shape descriptions, an interesting approach is proposed by the so-called view-based, or 2D/3D methods. In this case, instead of directly describing the 3D shape information, the 3D object is represented as a set of 2D images associated to the 3D object, corresponding to 3D/2D projections from several viewing angles. The 2D projection images are finally described by 2D descriptors. Here again, we distinguish two distinct approaches. A first one uses exclusively 2D information derived from the projections (*e.g.* resulting support regions, silhouette images…). The second one integrates some 3D information by considering the depth maps derived during the 3D/2D projection process [35, 36, 34].

A first and major advantage of such 2D/3D shape representations is essentially related to topological aspects: describing 2D images represented on a fixed topology (*i.e*. corresponding to a regular 2D lattice of pixels) is more tractable in practice than analysing 3D meshes with arbitrary and often non-regular connectivity. In addition, such methods offer high retrieval rates, quite competitive with most promising purely 3D approaches [21]. Finally, 2D/3D approaches that

exploit uniquely 2D features (*i.e*. no depth maps) open new perspectives in terms of applications. Notably, the main interest of such methods is related to the possibility of matching 3D models with 2D objects identified (*e.g*., with the help of some segmentation techniques) in still images or videos. In particular, such an approach might bring interesting and original solutions to the well-known problem of semantic gap [23, 24], which can be synthesized as follows. Given an object or a set of objects detected with the help of computer vision techniques from still images or videos, how can we interpret its meaning?

Existing approaches make intensively use of machine learning techniques [24], in order to bridge the gap between rough pixels and semantically meaningful interpretations of the image content. In this context, 3D modelling offers an interesting and complementary axis of research. Large and semantically categorized 3D repositories are today available. Thus, reliable 2D/3D matching techniques can lead to a semantic labelling of the content.

Let us also note that the advantage of using 3D models when searching for 2D images comes from the completeness of such representations. Thus, 2D representations of the same 3D object can present extremely different appearances in terms of shape. In this case, solely the 3D knowledge can relate effectively such different views (Figure 1).

**Figure 1:** Different views of a 3D object representing a bicycle. The first (profile) and the last (front) views are completely different in terms of 2D shape, but can be related if the 3D model is available.

In this paper we notably tackle the issue of shape-based 2D/3D indexing and propose a state of the art in the field. Solely 2D/3D techniques employing exclusively 2D shape information will be presented, since our objective is to investigate if such approaches can provide useful solutions for semantic identification of objects detected in 2D images/videos.

The rest of the paper is organized as follows. Section 2 proposes a state of the art of 2D/3D shape retrieval approaches. Generic principle and main families of methods are here described and discussed. The analysis of the literature shows that a key issue concerns the viewing angle selection procedure used to generate 2D projections. For this reason, in Section 3, we propose an experimental protocol and evaluate, under the same experimental conditions (*i.e*. test set, 2D shape descriptor and similarity metrics), the impact of the considered viewing angles on the performances of both 3D to 3D model retrieval and image recongnition from a single view applications. Finally, Section 4 concludes the paper and opens perspectives of future work.

## 2. STATE OF THE ART

Let us first briefly recall the principle of shape-based 2D/3D indexing techniques.

### 2.1. Shape-based 2D/3D indexing: principle

The underlying assumption of all 2D/3D methods can be stated as follows: two 3D objects are similar *iff* the associated 2D projections (with respect to a set of viewing angles) are similar.

Each 3D model $M$ is projected and rendered in 2D from $N$ different viewing direction $\{ n_i \}$. For simplicity, we will assume that the 3D object is centered in the origin of a 3D Cartesians system ($Oxyz$) and that the viewing angles $n_i$ are defined as samples on the unit sphere. A set of silhouette images (*i.e*. binary projections of the object), denoted by $\pi_i(M)$ is thus obtained (Figure 2). Each projection $\pi_i(M)$ is further described by a 2D shape descriptor $d_i(M)$.

When considering such an approach, some fundamental questions are coming out: how many projections should we choose? Which are the viewing angles that optimally represent the shape? Which are the appropriate descriptors/representations that can effectively describe the 2D shape of the obtained projections? How to minimize the computational complexity of the matching algorithms?

In the same time, when considering the issue of 2D/3D retrieval, invariance aspects with respect to similarity transforms (*i.e*. rotation, translation, isotropic scaling and combinations of them) should be taken into account. This issue is

particularly critical in the case of 2D/3D approaches: due to self-occlusions of object's subpart and global perspective deformations, the 2D shape of the corresponding silhouette images can change drastically from one view to another. The selection of an appropriate set of viewing angles is crucial for achieving a 3D pose invariance behavior.

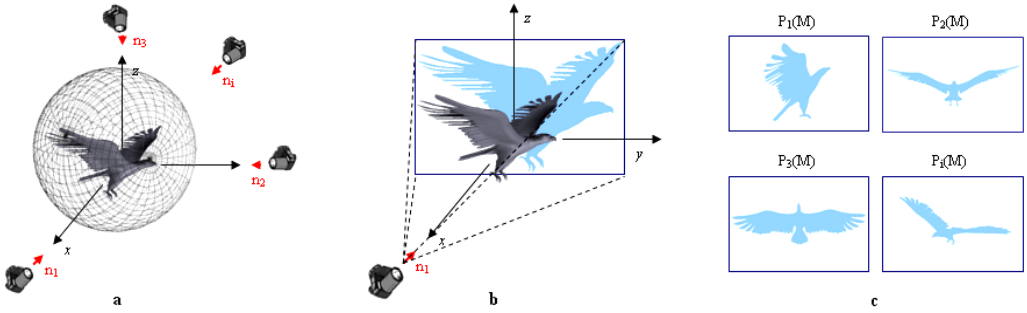



**Figure 2:** Projecting a model (model from NIST 3D model database): a. Viewing directions; b. Model projection according to the $n_1$ direction; c. the resulting silhouette images.

Since this issue is fundamental for successful 2D/3D similarity retrieval, we have categorized the various families of approaches with respect to the viewing angle selection procedure involved. A first and largely popular family of approaches considers a principal component analysis (PCA) of the 3D geometry, in order to obtain an object-dependent, canonical coordinate system invariant under 3D rotation. Such methods are presented in section 2.2. A second solution, further presented in section 2.3, consists of evenly distributing the camera viewing angles around the object.

### 2.2. Methods using PCA-based projection

Principal component analysis (PCA) is a well-known statistical procedure which optimally decorrelates a set of multi-valued variables. In our case, the variables to be considered represent the vertex positions of the 3D object expressed as vectors in a 3D coordinate system.

The object $M$ is first placed with the gravity centre in the origin of the considered coordinate system. Then, the (3×3) covariance matrix $C_M$ is computed, as described by the following equation:

$$C_M = \frac{1}{V}\sum_i v_i \, v_i^t \quad , \tag{1}$$

where $v_i = \left(v_i^x, v_i^x, v_i^x\right)^t$ is the 3D coordinate vector of the $v_i$ vertex, $V$ is the total number of vertices and subscript $t$ denotes the transpose operator.

As defined, the covariance matrix $C_M$ is symmetric and positive defined. Thus, it can be diagonalized by an orthogonal transform whose columns are its normalized, unit-length eigenvectors. The eigenvectors are also called *principal axes* or *axes of inertia* and the corresponding image planes (*i.e.* projection planes orthogonal to the principal axes) and referred to as *principal planes*.

In the case of PCA-based methods, the principal axes are used as viewing angles. They offer an object-dependent, orthogonal coordinate system. The orthogonality of the eigenvectors ensures the minimization of the redundancy between views. However, in order to obtain a more complete representation, additional viewing angles can be defined starting from the three eigenvectors. For example, the MPEG-7 *Multiview* description scheme (DS) [38] recommends the use of the diagonal directions of the four octants of the semi-space defined by the principal axes (Figure 3).

Let us also note that the corresponding eigenvalues provide a measure of the object's extent along each principal axis, which can be used for scale normalization [28].

As representative of the 2D/3D shape-based retrieval approaches, let us first mention the *Multiview DS* proposed by MPEG-7 standard. Within the MPEG-7 framework, two different MPEG-7 2D shape descriptors [31, 32, 37], the ART (*Angular Radial Transform*) [33] and CSS (*Contour Scale Space*) [6], can be considered. The MPEG-7 approach is described in details and experimentally evaluated in [21]. A maximum number of seven projection images can be used. Three of them

correspond to the principal directions and the other four to the diagonal, secondary views. The pre-processing stage involves translation and scaling, the 3D object being transformed such that its gravity centre coincides with the coordinate system origin and fits the unit sphere.

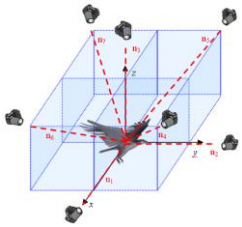


**Figure 3:** Selection of the principal and secondary axes.

In the case of the 2D-ART descriptor, the object's support function is represented as a weighted sum of 34 ART radial basis functions. In order to achieve rotation invariance, solely the absolute values of the coefficients are used. The similarity measure simply consists of $L_1$ distances between ART coefficients. The 2D-ART is thus invariant under similarity transforms, and is suitable for meshes of arbitrary topologies, which can present holes or multiple connected components under the projection operator. More restrictive, the MPEG-7 CSS descriptor assumes that the shape of the object can be described by a unique closed contour. The CSS descriptor is obtained by successively convolving an arc-length parametric representation of the curve with a Gaussian kernel. The curvature peaks are thus robustly determined in a multi-scale analysis process and serve to characterize the contour shape, with their value and corresponding position (expressed as curvilinear abscise).The associated similarity measure between two contour in CSS representation is based on a matching procedure which takes into account the cost of fitted and unfitted curvature peaks.

Whatever the 2D shape descriptor considered, when comparing two 3D models, a distance value $e_{ij}$ is obtained for each pair of *i* and *j* views. An error matrix $E=(e_{ij})$ is thus computed. The global similarity measure between the two models is then defined as:

$$D(M_1, M_2) = \min_{p \in \Pi} \{ Trace[\ p(E)\ ] \} \tag{2}$$

where $\Pi$ represents the set of all possible permutations between the columns of matrix *E*, *p* is a permutation in $\Pi$, and $p(E)$ represents the permuted version of matrix E.

Let us note that such a similarity measure is highly expensive since the number of possible permutations is non-polynomial with the number of views. In practice, such a similarity measure can be applied only for a reduced number of views and becomes computationally un-tractable when the number of views exceeds 7.

In [5], authors re-consider the MPEG-7 CSS representation. Here, the contour of each projection image is represented in CSS and decomposed into a set of segments called *tokens*, *i.e.* sets of 2D points delimited by two inflexion points. The tokens are then clustered and hierarchically organized in a M-Tree structure [25]. To compare two tokens, a sum of geodesic distances between points is computed. The obtained descriptor is intrinsically invariant to similarity transforms.

In [8] authors propose a 2D/3D approach based on the MCC (*Multi-scale Convexity/Concavity*) representation introduced in [26]. The 3D object is scaled with respect to its bounding sphere and PCA is applied in order to normalize the pose of the model. A number of three to nine silhouettes are then computed. As in the case of the MPEG-7 CSS descriptor, a scale-space analysis is here performed. Each silhouette contour is successively convolved with $K = 10$ Gaussian functions, with increasing kernel bandwidths. The variations (in 2D position) of 100 sample points, measured between successive scale spaces are then exploited as a shape descriptor. The similarity measure proposed involves a point to point matching procedure, which is quadratic with the number of sample points. In the same paper, a second method, so-called *Silhouette Intersection* (SI), is proposed. Only the three views, corresponding to the PCA principal directions are here exploited. The signature of a model is simply constituted by the three obtained binary silhouettes. The distance between two silhouette images is defined as the number of pixels belonging to the symmetric difference [27] of the two silhouettes (Figure 4). The global distance between two 3D objects is defined as the sum of silhouette distances between pairs of images corresponding to the same axes. However, the SI method is not robust against small variations of the shape. In addition, results strongly depend on the PCA alignment.

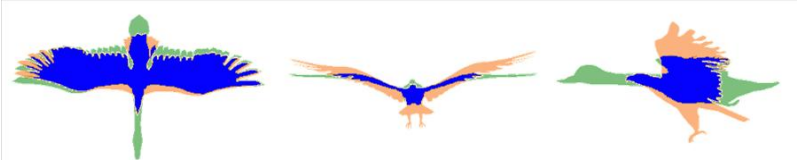
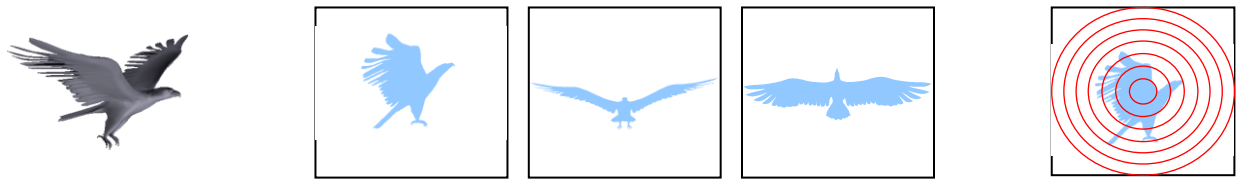


**Figure 4:** Example of two superposed silhouettes (here, a duck and an eagle). Even if the two models are similar, there is a large number of pixels belonging to the symmetric difference.

A different, single view approach is proposed recently in [4]. Authors propose to exploit an unique projection onto the principal plane of maximal eigenvalue. The projection is described using two descriptors: region-based Zernike moments [7] and contour-based Fourier descriptors [22]. The algorithm is very intuitive and fast, since a single projection image is used.

Finally, let us mention the approach proposed in [13]. A voxelized, volumetric representation is determined prior to performing the PCA. The viewing angle directions used to project the model correspond to the three principal axes. Each of the obtained images is decomposed into *L=60* concentric circles defined around the object's gravity center (Figure 5). The number of pixels located within each circle is computed, normalized to the total obects' and forms the feature vector associated to each projection. The so-called *principal plane descriptor* (PPD) proposed is defined as the set of all three feature vectors. Let us note that, because of the concentric circular regions involved, the PPD is intrinsically invariant under 2D rotation.



a. b. c.

**Figure 5:** Principle of the PPD approach: a. the 3D model; b. the object's projections onto the principal planes; c. concentric circles used to determine the descriptor.

However, the method implicitly assumes an ordering of the three principal directions (*i.e.* by decreasing values of corresponding eigenvalues). Such an approach may lead, in the general case, to miss-alignments, as shown in [20, 14]. This problem is illustrated in Figure 6.

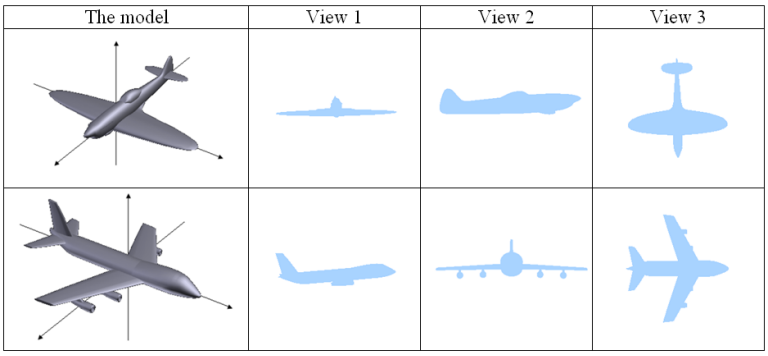



**Figure 6:** PCA miss-alignment. The choice of the principal axes is different for the two models with respect to their components which creates wrong matches (view 1 and view 2).

PCA-based methods offer the advantage of obtaining a representation associated with a canonical, object-dependent coordinate system that can partially solve the 3D transform invariance issues. However, the principal axes may present strong variations when dealing with similar models [20, 14]. Furthermore, a second limitation is related to the eventual miss-alignments that might occur. The reliability of the PCA is a key factoring this process that should be taken into account appropriately.

In order to overcome such limitations, a second family of methods, described in the next section, proposes to perform the 3D/2D projection according to a set of dense and evenly distributed viewing angles.

### 2.3. Methods using evenly distributed viewing angles

Instead of computing preferential projection planes, this second family of approaches uses a set of dense and evenly distributed viewing angles.

In [15] an extension of the SI algorithm [8], so-called *Enhanced Silhouette Intersection* (ESI), is introduced. Instead of exploiting the PCA directions, the vertices of a dodecahedron are here used to generate ten views, acquired after object normalization in translation, scale and rotation [30]. As in the case of the SI method, the distance between two images is given by the number of pixels included in the symmetric differences between the corresponding support regions. However, for the global distance between two 3D models, instead of summing up the distances between similar views, a weighted sum is proposed. This choice starts from the simple assumption that the relevance of a projection is proportional to the root square of its area. Both the descriptor extraction procedure and the dissimilarity measure are fast to compute. Compared with the SI algorithm, the robustness of the alignment is increased with the help of the approach described in [30]. However, articulated object matching is not supported and even small variations of the shape can greatly influence the retrieval results.

In [1], authors introduce the *LightField Descriptor* (LFD) which encodes ten silhouettes of the 3D object obtained by projection from the vertices of the dodecahedron. Translation and scaling invariance of the image are extrinsically achieved by normalizing the size of the projection images. Furthermore, the silhouettes are encoded by both Zernike moments [7] and Fourier descriptor [22]. A number of 35 coefficients Zernike moments are used and of 10 coefficients for the Fourier descriptor. Thus, the resulting descriptor includes 45 coefficients for each projection image, and 450 for each LFD associated with a 3D model. To compare two LFDs, the dissimilarity measured used is the $L_1$ distance between the descriptor's coefficients. The minimum sum of the distances between all possible permutations of views provides the dissimilarity between the 3D models. Let us note that in the case of LFD there are 60 possible permutations and for each of them 10 individual distances between pairs of images need to be computed. In addition, as one LFD is not totally invariant under rotation, a set of 10 LFDs per model is used to improve the robustness. This leads to a total number of 5460 comparisons to be computed. The need for multiple matches for each two objects makes LFD very time consuming. In order to reduce the computational cost, a multi-step fitting approach is adopted. In the first stage a reduced number of images per model and of coefficients is used in order to filter the results and retain a reduced number of candidate models. This procedure allows the early rejection of non-relevant models. The results obtained show that this algorithm outperforms most 3D shape descriptors at the cost of a significantly increased computational complexity.

An modified version of the LFD method is proposed in [17]. Authors start from the observations that two different objects can have similar projections (Figure 7), under the assumption that the scaling normalization is performed upon the silhouette images.

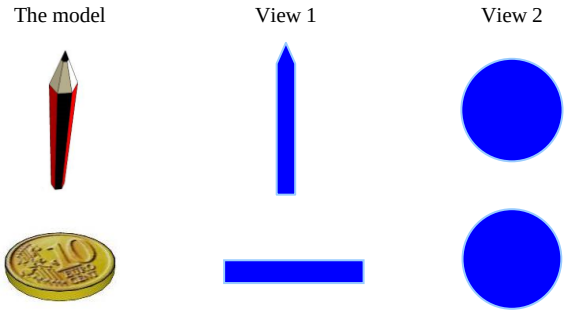



**Figure 7:** Dissimilar objects presenting similar views after scale normalization.

The Modified LFD (MLFD) approach proposed skips the resizing step of the original LFD algorithm. Tested on the Princeton Shape Database, the MLFD slightly outperformed the LFD performance (the Nearest Neighbor measure increases by 5.3%, the First and Second Tier measures by 4.1%, respectively 3.6% and the Discounted Cumulative Gain increases by 2.3%). This shows that the normalization issues, needed for achieving invariance, should be considered in the 3D space rather than in the domain of 2D projections and/or descriptors.

Recently, a Compact Multi-View Descriptor (CMVD) was proposed by [3]. The authors tested the CMVD on both binary and depth images. The descriptor extraction starts with the normalization stage, which includes translation, rotation and scaling. In order to compute the principal axes, both PCA and VCA [10] are performed. A number of 18 projections are obtained by placing the camera on the vertices of a 32-hedron and each of them is described by three sets of coefficients. First, 78 coefficients of the 2D Polar-Fourier Transform are computed. The usage of polar coordinates ensures rotation invariance. Secondly, 2D Zernike moments up to the $12^{th}$ order are obtained, resulting in 56 coefficients. Finally, 78 additional coefficients, corresponding to 2D Krawtchouk moments [19] are considered. When comparing two projections images, the $L_1$ norm is used to compute the distance between two descriptor vectors. Authors also take into account the fact that in some cases the first principal axis may not be successfully selected among the three principal axes. In order to deal with such an issue, $3 \times 8 = 24$ different alignments are considered. The total dissimilarity between two sets of images is obtained by summing up the dissimilarities between corresponding images. The distance between two models is the minimum distance that results when comparing the 18 projections of the first model with each of the 24 sets of images of the second model. In terms of computational complexity, the view generation process is the most time consuming. The 2D rotation invariance is ensured by the image descriptors considered. However, 3D rotation is provided only by the PCA&VCA alignment. Experiments were conducted on several 3D model databases and have shown that CMVD performs similar to LFD while offering a reduced computational complexity.

Some approaches further perform a selection of the resulting silhouettes in order to reduce their number. The selection process supposes to cluster the silhouettes according to their 2D similarity, resulting in a reduced number of representative projections that can be used to appropriately describe the 3D object.

In [2], authors present a method based on a similarity aspect graph. First introduced in [29], the aspect graph represents a subset of object projections obtained from a uniform viewing angle distribution. Here, the initial viewing angles are uniformly sampling, with a 5° step, the 3D object's principal plane in the (0, 180°) interval. A number of 36 silhouette images are thus obtained. A similarity metric is computed between each two successive projections. The *aspects* are defined as group of similar silhouettes with respect to the considered similarity measure. They are obtained with the help of a clustering algorithm that minimizes intra-class similarity while maximizing inter-class similarity. A *stable*, or *prototype* view is determined for each aspect. Finally, the prototypes views are represented as a graph structure where each node corresponds to a stable view (and each two adjacent stable views are connected by an edge of the graph.

Concerning the similarity matching process, only one projection image is used as query for each model, and compared to all the prototypes in the database. The number of comparisons is thus proportionally with the number of prototypes per model. The object is matched against the one in the database having the most similar prototypes to the query. The same similarity measure is used to compare prototypes as the one exploited for their selection. The results presented in the paper are obtained using the shock graph descriptor [12] which is time consuming in the matching stage.

The main drawback of this method is the computational complexity of both prototype selection and similarity measure used for retrieval purposes. Also, since the viewing angles lie in a unique plane, the selection of this plane has to generate a robust response in order to ensure a 3D pose invariant behaviour.

In [16] a similar method is proposed. The approach aims at selecting the viewing angles based on the degree of representativity of the corresponding projection images. Here, 162 silhouette images are rendered according to a uniform viewing angle distribution, defined over the unit sphere as a spherical triangular mesh. The obtained images are described using Zernike moments [7] up to order 15. The similarity between each two adjacent projections is computed using the $L_2$ norm distance. Then, a spherical weighted graph is constructed using the viewing angles as vertices. The weight of each edge is equal to the similarity between the views connected by the considered edge. Stable view regions represent sub-parts of the graph with similar corresponding projections (*i.e*. sets of edges with low weights). Two stable view regions are separated by so-called *heavy edges*. Thus, based on the edge weights, the graph is partitioned into eight sub-graphs representing the stable view regions. Furthermore, a representative viewpoint needs to be found for each stable region. In order to achieve this goal, a pertinence value is assigned to each viewpoint. This value is based on the mesh saliency, a measure that evaluates the mesh curvature evolution when smoothed at different filter scales. The saliency measure is also used to sort the views according to the amount of information they carry. This algorithm is complex because of the weighted graph construction which is the most time consuming stage (162 vertices of 6 adjacent edges each generate 486 edges and as many weights to be computed).

The proposed algorithm proves to be effective for determining representative viewpoints. However, it does not address the issue of 3D model matching. Notably, it would be interesting to establish in what measure the viewing angles thus determined could be effectively exploited within the framework of 2D/3D shape-based retrieval applications.

In general, the methods that use evenly distributed viewing angles generate a higher number of projections (100 views for LFD, 18 views for CMVD) than those based on the PCA analysis (between 3 and 9 views) Obviously, a larger number of images will carry more information which results in a more complete description. In the case of aspect graph and stable views algorithms, the redundancy is reduced by selecting a sub-set of representative images. However, the main drawback of such approaches remains their high computational complexity.

Table 1 synthesizes the various descriptors presented in this section. For each method, the extraction and the matching complexities, respectively denoted by $C_E$ and $C_M$, are qualitatively estimated (*i.e.* + for low complexity and +++ for high). The numbers of views per model as well as the viewing angle selection procedure are also indicated. The last column recalls the 2D descriptor used to describe the projection images.

**Table 1.** Overview of 2D/3D approaches ($C_E$ – Descriptor's extraction complexity, $C_M$ – Matching complexity).

| Method | $C_E$ | $C_M$ | No of views | Viewing angle selection | 2D descriptor |
|---|---|---|---|---|---|
| **PPA** [4] | + | + | 1 | PCA | Zernike moments & contour-based Fourier descriptor |
| **PPD** [13] | + | + | 3 | PCA | Sums of pixels into concentric circles |
| **MPEG-7 2D/3D CSS** [37] | + | ++ | 7 | PCA | Curvature scale space Descriptor |
| **MPEG-7 2D/3D ART** [37] | ++ | ++ | 7 | PCA | Angular Radial Transform |
| **MCC** [8] | ++ | +++ | 3/9 | PCA | Curve evolution when filtered with Gaussians |
| **SI** [8] | + | + | 3 | PCA | Binary images |
| **Aspect graph** [2] | +++ | +++ | 5 – 10 | Aspect graph prototypes | Shock graph |
| **Stable views** [16] | +++ | N/A | 8 | Spherical graph stable views | Zernike moments coefficients (up to order 15) |
| **ESI** [15] | + | + | 10 | Even distribution | Binary images |
| **LFD** [1] | +++ | +++ | 100 | Even distribution | Zernike moments & Fourier descriptor |
| **CMVD** [3] | ++ | ++ | 18 | Even distribution | Zernike moments, Polar Fourier and Krawtchouk moments coefficients. |

The literature shows a wide palette of useful approaches. However, evaluating the importance of the viewing angle selection process remains difficult, since methods use different descriptors and normalization procedures. In order to perform a fair comparison, we have established an experimental evaluation protocol, described in the next section.

## 3. EXPERIMENTAL EVALUATION

Experiments have been carried out on the MPEG-7 test dataset [21], which includes 363 models semantically categorized in 23 classes. Categories include humanoids, airplanes, cars, trees (with or without leafs), five synthetic letter models (from A to E), rifles, missiles, pistols, helicopters, motorcycles… Some sample models are illustrated in Figure 8. The categories exhibit a relatively important intra-class variability [20] in terms of 3D shape. This ground truth databases allows us to perform an objective comparison of the different approaches.

Concerning the viewing angles considered, we have considered both PCA and uniform distributions.

For the PCA approach, a number of 3 and 7 directions, corresponding to the three principal and four secondary axes, have been retained (*cf.* Section 2.2).

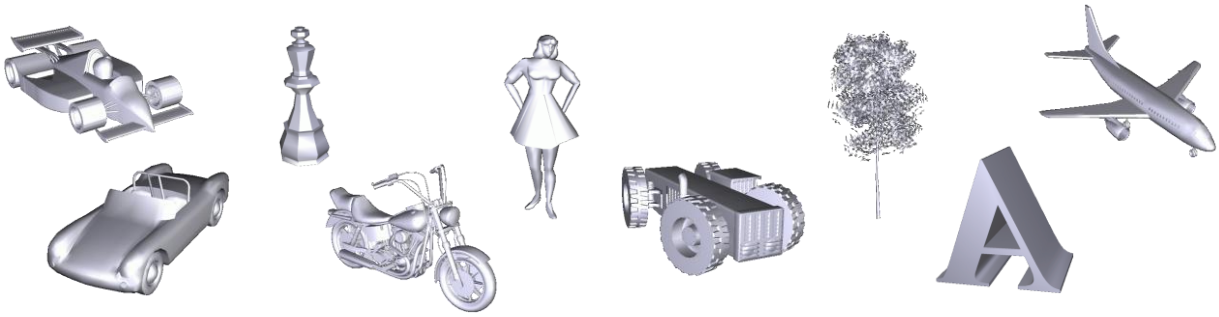
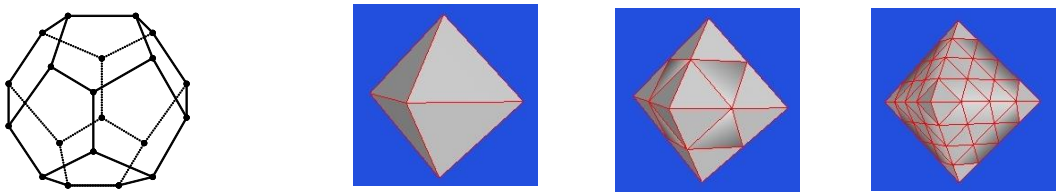


**Figure 8:** Sample models from the MPEG-7 3D dataset.

Concerning the even distributions, we have first adopted the dodecahedron-based strategy proposed by the LFD approach, which leads to 10 viewing angles (Figure 9.a). In addition, we have considered a second strategy, based on successive subdivisions of a regular octahedron [20, 21] (Figure 9.b). The initial octahedron gives 3 distinct axes, corresponding to 3 out of 6 vertices. At one level of subdivision, 9 viewing angles are obtained. Finally, at two levels of subdivision, 33 directions can be used.



**a.** Dodecahedron

**b.** Recursive subdivisions of a regular octahedron.

**Figure 9:** Uniform viewing angle selection with dodecahedron (LFD) and octahedron approaches.

Whatever the viewing angle selection strategy, a same descriptor is associated to each projection image. For the sake of generality, we have adopted the MPEG-7 2D ART descriptor with 35 coefficients, which is able to deal with silhouettes of arbitrary topologies.

As objective evaluation measures, we have retained the *First Tear* (FT) and *Second Tear* (ST) Scores [9], previously used in [21], which respectively represent the percentage of correct retrieved models within the Q and 2Q first retrieved results (where Q denotes the total number of models of the query's category).

In a first time, we have considered the issue of 3D to 3D model retrieval.

### 3.1. 3D to 3D model retrieval

Two similarity metrics have been adopted. The first one, so-called *diagonal*, assumes that the PCA alignment is correct for all models in the database. Thus the similarity between two 3D models is computed as the sum of distances between 2D ART descriptors associated to the corresponding 2D silhouettes. The second one, so-called *minimum*, exploits a greedy strategy for fitting the various 2D views. When comparing two 3D objects, the best match, corresponding to the minimal distance between all combinations of views is first determined. The corresponding views are considered as aligned and the process is successively applied upon the remaining sets of views, until all the views are matched.

Concerning the normalization issues, each object has been first centered in the origin of the 3D coordinated system. A PCA analysis has been applied in order to align the object's principal axes with the systems axes. The PCA eiegenvalues, which provide a measure of the object's size have been exploited for scale normalization, as described in [20]. In order to evaluate the impact of the PCA alignment, we have also considered the following experiment, applied in the case of the LFD strategy. After centering and scaling, the object has been rotated with a random 3D rotation matrix. In this way, we are able to investigate the retrieval performances when the object has an arbitrary pose.

Table 2 presents the FT and ST results obtained with different viewing angles and similarity metrics. The LFD column states for applying the LFD projection on the randomly rotated object, while LFDPCA concerns a preliminary alignment of the object with the PCA axes.

**Table 2:** Retrieval FT and ST scores (%) obtained on the MPEG-7 dataset (in bold, the maximum scores obtained).

| Matching strategy | Score | PCA3 | PCA7 | LFDPCA | LFD | OCTA9 | OCTA33 |
|---|---|---|---|---|---|---|---|
| **Minimum** | **FT** | 63.65 | 64.10 | 63.72 | 59.87 | **65.44** | 65.21 |
| **Diagonal** | | 65.57 | 66.11 | 65.37 | 45.19 | 66.57 | **67.47** |
| **Minimal** | **BE** | 71.86 | 73.15 | 71.61 | 70.66 | 73.73 | **73.84** |
| **Diagonal** | | 73.76 | 73.54 | 73.50 | 58.77 | **75.56** | 74.71 |

When considering the *Minimum* matching metric, the maximal retrieval scores are achieved for the octahedron-based viewing angle selection strategy with 9 (OCT9) and 33 (OCT33 views, with FT and ST scores up to 65.44% and 73,84%). However, such scores are very lightly superior to those obtained with PCA3 (FT = 63,65%, ST = 71.86%) and PCA7 (FT = 64,10%, ST = 73.15%). This shows that increasing the number of views does not necessarily leads to a spectacular gain in term of retrieval efficiency. This can be explained by the fact that in this case, the number of false positives responses provided by the 2D ART also increases with the number of views. In addition, in the cases of objects presenting symmetries, several of the views considered can be similar, which can further bias the results.

In the case of the *Diagonal* matching strategy, the scores are slightly superior, with an average gain of about 2%.

Finally, when comparing LFD and LFDPCA approaches, we can observe that the scores are improved by the PCA alignement. This is natural when considering the *Diagonal* matching, where we obtain a significant gain (20%), but holds also in the case of the *Minimal* strategy (with a gain of 4%), and thus shows the relevance of the PCA alignment.

A second experiment concerns the aspects of image recognition from a single view.

### 3.2. Image recognition from a single view

A test set has been first constituted. It includes 46 3D objects, randomly selected from the MPEG-7 dataset, such that each category is represented by two objects. For each test object, different test images have been generated. A first one gives a frontal view and corresponds to the object's projection onto the principal plane (defined by the principal axes of first and second largest eigenvalues). A second image corresponds to the projection from a randomly generated angle of view. Finally, a third one, is also randomly generated but in this case the angle of view is restricted to a ±45° interval around the normal to the principal plane. This is motivated by the fact that in videos, objects pose variation can be considered as limited.

Two performance metrics have been retained in this case. The first one, so-call *First Answer* (FA) provides the percentage of queries where an image from the correct category has been found on the first retrieved position. The second one is the *category recognition rate* (CRR), defined as follows. For each query, the categories obtained in the first 10 retrieved images are determined. We decide then that the query object belongs to the most represented category among them (*i.e*. category with the greatest number of objects retrieved in the top 10 results). The CRR is finally defined as the percentage of correct decisions over the whole test set of 46 models.

Table 3 presents the obtained results.

In the case of the randomly generated query images, a set of 10 images has been generated for each type of random trial (*i.e.* RAND and RAND45). The results reported in Table 3 represent the average scores obtained.

The best FA and CRR scores are obtained when using the PP image (FA = 82.6% and CRR = 69.56, for OCTA33 representations). This shows the high relevance of using the frontal image in the retrieval process. However, in the case of real objects detected from images or videos, we cannot ensure that the PP image can always be detected.

When using the completely random images (RAND), the recognition scores are drastically lower. In this case, the best FA score is achieved by OCTA33 (67,39%). Concerning the CRR, the maximum value is obtained by the LFD representation (56,62%). The recognition rates are increasing when limiting the 3D object's pose variation in the case of RAND45. Here again, the best FA is provided by OCTA33 (67,39%), while the maximum CRR is given by the LFD approach (67,78%). Let us note that this score approaches the optimal obtained for the PP image with OCTA33 (69,56%).

**Table 3:** Image recognition from a single view with FA and CRR scores in function of the different viewing angle selection strategies and of the type of query image (PP – frontal view on the principal plane, RAND – completely random pose, RAND45 – random pose restricted to a ±45° around the normal to the principal plane). The best scores are represented in bold.

| Query image type | Score | PCA3 | PCA7 | LFD | LFDPCA | OCTA9 | OCTA33 |
|---|---|---|---|---|---|---|---|
| **PP** | **FA** | 71.73 | 76.08 | 78.26 | 54.34 | 80.43 | **82.60** |
| **RAND45** | | 45.65 | 52.17 | 69.56 | 60.86 | 56.52 | **73.90** |
| **RAND** | | 21.73 | 47.82 | 65.21 | 56.52 | 47.82 | **67.39** |
| **PP** | **CRR** | 63.04 | 63.04 | 63.04 | 47.82 | **69.56** | **69.56** |
| **RAND45** | | 39.32 | 52.56 | **67.78** | 58.10 | 58.89 | 66.79 |
| **RAND** | | 23.91 | 45.65 | **56.52** | 52.17 | 43.47 | 50.00 |

The obtained results clearly show the superiority of strategies based on evenly distributed viewing angles with respect to PCA-based approaches (*e.g*. with gain of 10% to 15% for LFD w.r.t. to PCA7), for image recognition purposes. In addition, the perfectly even distribution associated to the LFD approach offers highly interesting performances.

## 4. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an overview of the state of the art 3D/3D shape-based retrieval approaches. Two great families of approaches have been identified and described. The first one concerns methods based on a PCA alignment, while the second is based on dense and evenly distributed viewing angles. Uniquely approaches that are using exclusively 2D features have been considered.

In order to study the influence of viewing angles selection, we set up an experimental evaluation protocol, related to applications of both 3D to 3D model retrieval and image recognition from a single view. As 2D shape descriptors, we have considered the MPEG-7 2D ART. Experimental results, obtained on the MPEG-7 ground truth dataset show that:

- for 3D to 3D model retrieval, increasing the number of views do not significantly enhance the retrieval performances, due to false positives and miss-alignments,
- in the case of image recognition the evenly distributed viewing angles strategies lead to superior performances.

Our future work concerns the extension of this study to other, more discriminant descriptors, such as MPEG-7 CSS, Hough 2D, Fourier descriptors, or combination which can provide complementary 2D shape information for a more complete characterization of the 2D shape that can further increase the recognition rates. At a longer term, our work will concern with real-life objects, detected from natural videos and still images.

## 5. ACNOWLEDGMENT

This work has been partially supported by the UBIMEDAI Research Lab, between Institut TELECOM and Alcatel-Lucent Bell-Labs.

## REFERENCES

1. Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen and Ming Ouhyoung, "On visual similarity based 3D model retrieval", *Computer Graphics Forum*, vol. 22, no. 3, pp. 223-232, 2003.
2. C. Cyr and B. Kimia, "3D object recognition using shape similarity-based aspect graph", *Proc. 8th IEEE Int. Conf. Comput. Vision*, Vancouver, BC, Canada, pp. 254–261, 2001.
3. Petros Daras, Apostolos Axenopoulos, "A Compact Multi-View descriptor for 3D Object Retrieval", *International Workshop on Content-Based Multimedia Indexing*, June 2009.
4. YuJie Liu, Xiao-Dong Zhang, ZongMin Li; Hua Li, "3D model feature extraction method based on the projection of principle plane", *Computer-Aided Design and Computer Graphics, 2009 (CAD/Graphics '09)*, Page(s):463 – 469, August 2009.
5. S. Mahmoudi and M. Daoudi, "3D models retrieval by using characteristic views", *Proceedings of the 16th International Conference on Pattern Recognition*, pp. 457-460, Quebec, Canada, 2002.
6. F. Mokhtarian, A.K. Mackworth, "A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, pp. 789-805, August 1992
7. R. Mukundan and K. R. Ramakrishnan, "Moment Functions in Image Analysis: Theory and Applications", *World Scientific Publishing Co Pte Ltd.*, September 1998.

8. T. Napoléon, T. Adamek, F. Schmitt,N.E. O'Connor, "Multi-view 3D retrieval using silhouette intersection and multi-scale contour representation", *SHREC 2007 - Shape Retrieval Contest*, Lyon, France, June 2007.
9. R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions", *ACM Transactions on Graphics (TOG)*, vol. 21, pp. 807-832, October 2002.
10. J. Pu, and K. Ramani, "An Approach to Drawing-Like View Generation From 3D Models", *In Proc. Of International Design Engineering Technical Conferences*, September 2005.
11. Zheng Qin, Ji Jia, Jun Qin, "Content Based 3D Model Retrieval: A survey", *Content Based Multimedia Indexing 2008*, June 2008.
12. T. B. Sebastian, P. N. Klein, and B. B. Kimia, "Alignment based recognition of shape outlines", *Proceedings of the 4th International Workshop on Visual Form*, May 2001.
13. J.L. Shih, W.C. Wang, "A 3D Model Retrieval Approach based on The Principal Plane Descriptor", *Proceedings of The Second Internat. Conf. on Innovative Computing, Information and Control (ICICIC)*, pp. 59-62, 2007.
14. Johan W.H. Tangelder and Remco C. Veltkamp, "A Survey of Content Based 3D Shape Retrieval Methods", *Proceedings of the Shape Modeling International 2004 (SMI'04)*, pp. 145-156, Genova, Italy, 2004
15. Liuying Wen, Guoxin Tan, "Enhanced 3D Shape Retrieval Using View-Based Silhouette Representation", International Conference on Audio, Language and Image Processing, August 2008.
16. H. Yamauchi, W. Saleem, S. Yoshizawa, Z. Karni, A. Belyaev, H.-P. Seidel, "Towards Stable and Salient Multi-View Representation of 3D Shapes", *IEEE Int. Conf. on Shape Modeling and Applications, 2006*, pp.40-40, 14-16, June 2006.
17. Tao Yang, Bo Liu, Hongbin Zhang, "3D model retrieval based on exact visual similarity", *9th International Conference on Signal Processing*, December 2008.
18. Y. Yang, H. Lin, Y. Zhang, "Content-based 3D Model Retrieval: A Survey", *IEEE Trans/ on Systems, Man, and Cybernetics – Part C: Applications and Reviews*, Vol. 37, No6, p. 1081-1098, November 2007.
19. P.T.Yap, R.Paramesran and S.H.Ong, "Image Analysis by Krawtchouk Moments", *IEEE Transactions on Image Processing*, Vol. 12, No. 11, pp. 1367-1377, November 2003.
20. T. Zaharia, F. Prêteux, "3D shape-based retrieval within the MPEG-7 framework", *Proc. SPIE Conf. on Nonlinear Image Processing and Pattern Analysis XII*, Vol. 4304, pp.133-145, San Jose, CA, USA, January 2001
21. T. Zaharia, F. Prêteux, "3D versus 2D/3D Shape Descriptors: A Comparative study", *In SPIE Conf. on Image Processing: Algorithms and Systems*, Vol. 2004 , Toulouse, France, January 2004.
22. D. S. Zhang and G. Lu. "An Integrated Approach to Shape Based Image Retrieval", *Proc. of 5th Asian Conferenceon Computer Vision (ACCV)*, pp. 652-657, Melbourne, Australia, January 2002.
23. Tianyang Lv, Guobao Liu, Shao-bin Huang, Zheng-xuan Wang, "Semantic 3D Model Retrieval Based on Semantic Tree and Shape Feature", *Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 452-457, December 2009.
24. Boyong Gao, Herong Zheng, Sanyuan Zhang, "An Overview of Semantics Processing in Content-Based 3D Model Retrieval", *Artificial Intelligencw and computational Intelligence*, pp. 54-59, January 2010.
25. Ciaccia Patella Rabitti , P. Ciaccia , M. Patella , F. Rabitti , P. Zezula, "Indexing Metric Spaces with M-tree", *in SEBD 1997*, pp. 67-86, 1997.
26. T. Adamek and N. E. O'Connor, "A multiscale representation method for nonrigid shapes with a single closed contour", *IEEE Trans. Circuits Syst. Video Techn*, Volume 14, Issue 5, pp. 742–753, May 2004.
27. Helmut Alt, Ulrich Fuchs, Günter Rote, Gerald Weber, "Matching Convex Shapes with Respect to the Symmetric Difference", *Lecture Notes in Computer Science*, Volume 1136/1996, pp. 320-333, May 1998.
28. T.Zaharia, F. Preteux, "Shape-based retrieval of 3D mesh models", *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on Volume 1*, pp. 437-440, August 2002.
29. J. J. Koenderink, A. J. van Doorn, "The singularilarities of the visual mapping". *Biol. Cyber.*, Volume 24, pp. 51–59, 1976.
30. Chen Ding-Yun, Ouhyoung Ming, "A 3D model alignment and retrieval system", *Proceedings of International Computer Symposium, Workshop on Multimedia Technologies*, Hualien,Taiwan, pp. 1436-1443, December 2002
31. M. Bober, "MPEG-7 Visual Shape Descriptors", *IEEE Transaction on Circuits and Systems for Video Technology*, Volume 11, Issue 6, pp. 716-719, August 2002
32. B.S. Manjunath, Phillipe Salembier, Thomas Sikora, "Introduction to MPEG-7: Multimedia Content Description Interface", John Wiley & Sons, Inc., New York, NY, 2002.
33. W.-Y. Kim, Y.-S. Kim, "A New Region-Based Shape Descriptor", ISO/IEC MPEG99/M5472, Maui, Hawaii, December 1999.
34. J.L. Shih, C.H. Lee, and J.T. Wang, "A 3D Model Retrieval System Using the Derivative Elevation and 3D-ART"**,** *Proceedings of the 2008 IEEE Asia-Pacific Services Computing Conference*, p.p. 739-744, 2008.
35. R. Ohbuchi, K. Osada, T. Furuya, T. Banno, "Salient local visual features for shape-based 3D model retrieval", *Shape Modeling and Applications, 2008. SMI 2008*, pp. 93-102, June 2008.
36. P. Papadakis, I. Pratikakis, T. Theoharis, G. Passalis, and S. Perantonis, "3D object retrieval using an efficient and compact hybrid shape descriptor", *Eurographics Workshop on 3D Object Retrieval*, Crete, Greece, April 15, 2008
37. ISO/IEC 15938-3: 2002, MPEG-7-Visual, Information Technology – Multimedia content description interface – Part 3: Visual, 2002.
38. ISO/ IEC 15938-5, Information technology - MultimediaContent Description. Interface - Part 5: Multimedia Description Schemes. 2003.